SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

August 2015 Vol.58 No.8: 778–786
doi: 10.1007/s11427-015-4877-5

High-throughput sequencing-based genome-wide identification of microRNAs expressed in developing cotton seeds

WANG YanMei, DING Yan, YU DingWei, XUE Wei & LIU JinYuan*

Laboratory of Plant Molecular Biology, Center for Plant Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

Received August 8, 2014; accepted November 20, 2014; published online June 25, 2015

MicroRNAs (miRNAs) have been shown to play critical regulatory roles in gene expression in cotton. Although a large number of miRNAs have been identified in cotton fibers, the functions of miRNAs in seed development remain unexplored. In this study, a small RNA library was constructed from cotton seeds sampled at 15 days post-anthesis (DPA) and was subjected to high-throughput sequencing. A total of 95 known miRNAs were detected to be expressed in cotton seeds. The expression pattern of these identified miRNAs was profiled and 48 known miRNAs were differentially expressed between cotton seeds and fibers at 15 DPA. In addition, 23 novel miRNA candidates were identified in 15-DPA seeds. Putative targets for 21 novel and 87 known miRNAs were successfully predicted and 900 expressed sequence tag (EST) sequences were proposed to be candidate target genes, which are involved in various metabolic and biological processes, suggesting a complex regulatory network in developing cotton seeds. Furthermore, miRNA-mediated cleavage of three important transcripts *in vivo* was validated by RLM-5′ RACE. This study is the first to show the regulatory network of miRNAs that are involved in developing cotton seeds and provides a foundation for future studies on the specific functions of these miRNAs in seed development.

***Gossypium hirsutum*, seed development, microRNA (miRNA), target gene, GO annotation, high-throughput sequencing**

Citation: Wang YM, Ding Y, Yu DW, Xue W, Liu JY. High-throughput sequencing-based genome-wide identification of microRNAs expressed in developing cotton seeds. Sci China Life Sci, 2015, 58: 778−786, doi: 10.1007/s11427-015-4877-5

The unique feature of the seed of the upland cotton (*Gossypium hirsutum* L.) is that approximately 30% of the seed epidermal cells develop into specialized fibers while its embryo produces the high-quality oil and proteins [1]. The mature fiber is the world's most important natural textile fiber, whereas the high-quality oil and proteins can further increase the agronomic and economic importance of cotton as a crop plant. Although the growth process of cotton fiber has been intensively studied [2–4], as an integral part of the seed, increasing evidence shows that fiber growth is indeed regulated by seed development [5]. Therefore, it is necessary to have an in-depth investigation into the regulatory molecules of cotton seed development.

Cotton microRNAs (miRNAs) are endogenous, non-coding small RNAs that negatively control gene expression by degradation of target mRNAs [6]. Numerous studies have investigated cotton miRNAs and their regulatory functions in developmental processes of cotton fibers [6–10]. Especially, the high throughput RNA-sequence technology was employed to identify 562 candidate miRNAs from 7-DPA fibers of allotetraploid cotton [9], while cotton DD-genome shotgun sequences were examined to annotate 348 miRNA candidate sequences [2]. In our previous studies, seven fiber initiation-related miRNAs were characterized by comparative miRNAome analysis in developing cotton ovules [3]. Subsequently, from four small RNA libraries constructed from upland cotton fibers ranging from 5- to 20-DPA, eight miRNAs were found to be related to cotton fiber elongation [4]. Recently, 21 novel miRNA candidates were found to be expressed in secondary wall

*Corresponding author (email: liujy@mail.tsinghua.edu.cn)

thickening fiber cells [10]. These studies demonstrate potential regulatory functions for these small RNAs during initiation, elongation and secondary wall thickening processes of cotton fibers. However, no study to date has validated the distribution of cotton miRNAs and their expression in developing cotton seeds.

In the present study, to retrieve the firsthand information about the miRNA expressed in young cotton seeds, a small RNA library was constructed from 15-DPA seeds and subjected to high-throughput sequencing. As a result, more than 18 million short reads were produced and 95 known miRNAs and 23 novel miRNA candidates were observed to be expressed in cotton seeds. Moreover, putative targets for 21 novel and 87 known miRNAs were predicted. Our study provides a foundation for further investigation on specific functions of these miRNAs in cotton seed development.

1 Materials and methods

1.1 Plant material preparation and total RNA isolation

Upland cotton (*Gossypium hirsutum* cv. CRI35) was field grown under normal agronomic conditions. The seeds were kindly provided by the Cotton Research Institute at the Chinese Academy of Agricultural Sciences. Flowers were tagged on the day of anthesis. Seeds from cotton at 15 DPA were collected, immediately frozen in liquid nitrogen and stored at −80°C. The PureLink^{TM} Plant RNA Reagent kit (Invitrogen, USA) was used to extract total RNA from immature seeds at 15 DPA according to the manufacturer's instructions. The quality of the RNA was checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

1.2 Small RNA library preparation and high-throughput sequencing

Small RNA library construction was performed as described previously [3]. Briefly, 15–30 nt small RNAs were gel-purified from 15% PAGE (7 mol L^{-1} urea), 5′ and 3′ RNA adaptors were added, and RT-PCR using primers with partial complementarity to the adaptors was performed. The DNA pool was amplified from the first-strand cDNA and was then sequenced using Hiseq2000 (Illumina, USA) at the Beijing Genomics Institute (BGI), Shenzhen.

1.3 Identification and analysis of known and novel miRNAs

The raw sequences from Illumina Hiseq2000 were processed using the SOAPnuke software (http://soap.genomics.org.cn/) to filter the low-quality reads and adaptor sequences. Then small RNA reads matching non-coding rRNA, tRNA, snRNA and snoRNA in Rfam 12.0 database (http://rfam.xfam.org) were removed. The generated high-quality small RNA reads ranging from 18 to 30 nt were aligned against the *Gossypium raimondii* genome shotgun-sequence assemblies (http://cgp.genomics.org.cn/page/species/index.jsp) and those mapped to the cotton genome sequence assemblies were retained for further analysis. To identify previously known miRNAs, these small RNA sequences were subjected to a BLASTN analysis against miRBase (release 21: June 2014, http://microrna.sanger.ac.uk). The sequences identical or related to known miRNA sequences (with two or fewer nucleotide substitutions) were considered known miRNAs. The rest mapped sequences were analyzed for predictions to identify novel miRNA candidates by the mireap_0.2 program (http://sourceforge.net/projects/mireap). Briefly, the cotton genome was used as reference to explore the potential precursors for novel miRNA candidates in *G. raimondii*, and the obtained precursor sequences were examined for the potential to form secondary structures. The secondary structures were further checked for free energy, dominance of the novel miRNA reads relative to other precursor-mapped small RNA reads in abundance, the number of mismatches between the miRNA and the other arm of the hairpin, and no more than two asymmetric bulges in the stem region, to meet the gold criteria defined previously for annotation of plant miRNAs [11]. We calculated the minimal folding free energy index (MFEI) of all candidates to confirm whether they were true miRNAs. Based on previous studies, the value of MFEI for the most potential precursors was greater than 0.85, which is remarkably higher than that of other non-coding small RNAs. The MFEI was calculated as MFE/(precursor length)×100/(G+C)% [12].

The miRNA data from the 15 DPA small RNA libraries in our previous study were used as a reference to gain insight into the miRNA expression profiles [4]. The relative abundances of known and novel miRNAs in these two libraries were reported as normalized reads (reads per ten million, RPTM). The Student's *t*-test was used to assess the statistical significance of the differences between the highest and lowest abundance levels of each miRNA for miRNA families with reads greater than 100 RPTM, as previously described [3]. Secondary structures were predicted with MFOLD (http://mfold.rna.albany.edu) using sequences identified by mireap.

1.4 Target gene prediction and GO annotation

The potential target genes of cotton miRNAs were searched against the *Gossypium* (cotton) DFCI Gene Index (CGI) (release 11, http://compbio.dfci.harvard.edu/index.html) using the web-based tool psRNATarget program (http://plantgrn.noble.org/psRNATarget/). The most abundant miRNA variants were served as queries. The default parameters were used as follows: maximum mismatches between the miRNA and target: 3; multiplicity of target sites: 2, and no mismatches within the maximum expecta-

tion region. To better annotate the potential target function, a BLASTN search was carried out against database at the NCBI, followed by Gene Ontology (GO) enrichment analysis using the AgriGO web service (http://bioinfo. zcau.edu. cn/agriGO/index.php) [13].

1.5 RLM-5′ RACE validation of miRNA targets

To validate the cleavage sites of target genes, a RNA ligase-mediated rapid amplification of 5′ cDNA ends (RLM-5′ RACE) was performed using the First Choice RLM-RACE kit (Ambion, USA). Total RNA was extracted with PureLink^{TM} Plant RNA Reagent (Invitrogen) from the immature seeds at 15 DPA, and poly(A)$^+$ mRNA was purified with an mRNA Purification Kit (Invitrogen). The purified poly(A)$^+$ mRNA was ligated to the 45 nucleotide adapter from the kit. Subsequent steps were followed according to the manufacturer's instructions. The PCR-amplified products were separated and cloned into pEASY-T1 vector (TransGen, Beijing) for sequencing. All gene-specific primers used in the experiments are listed in Table S1.

2 Results and discussion

2.1 Overview of deep sequencing datasets

To examine the expression of miRNA genes in developing cotton seeds, a small RNA library was generated from immature seeds at 15 DPA and sequenced by an Illumina HiSeq 2000 analyzer. After removing low-quality reads and adaptor sequences, a total of 18,562,708 reads representing 7,161,027 unique reads was obtained from the library (Table 1). These clean reads were further mapped to the diploid cotton *G. raimondii* genome [2], generating 5,612,820 genome-matched reads (30.23% of the clean reads). Approximately 0.41% of the unique reads matched non-coding RNAs including rRNA (0.32%), tRNA (0.05%), snRNA (0.01%) and snoRNA (0.01%), which comprised 5.58% of all sequenced reads (Table 1). The majority of the reads (88.71%) did not match known small RNAs and possibly represent new regulatory small RNAs and novel miRNAs. However, only 25.62% of the unique reads were mapped to the diploid DD genome, suggesting that the remaining 74.38% of the unique reads might be from another allotetraploid cotton AA genome or from evolved AADD genomes.

The redundant and unique reads greater than 18 nt long in immature seeds are shown in Figure 1. Consistent with the length distribution pattern of small RNAs in other plant species, the majority of the obtained small RNA sequences were 20–24 nt in this study. The 24 nt small RNAs were the most abundant, representing 60%, which was a feature of some of the siRNAs. The next larger fractions were the 23 nt (17.3%), 21 nt (7%) and 22 nt (6%) fractions, representing the typical length of mature plant miRNAs. Generally, small RNAs in the 21 nt class represent miRNAs. The same phenomenon was also observed in peanut [14], soybean [15], maize [16], rice [17,18], and *Medicago truncatula* [19].

Table 1 Summary of the small RNA sequencing data

Category	Total reads (Percentage)	Unique reads (Percentage)
Clean reads	18,562,708 (100%)	7,161,027 (100%)
Genome matches[a)]	5,612,820 (30.23%)	1,834,912 (25.62%)
Known miRNA[b)]	773,117 (4.16%)	809 (0.02%)
rRNA[a)]	748,221 (4.03%)	23,118 (0.32%)
tRNA[a)]	283,468 (1.53%)	3,364 (0.05%)
snRNA[a)]	3,892 (0.02%)	949 (0.01%)
snoRNA[a)]	744 (0.00%)	949 (0.01%)
mRNA[a)]	196,748 (1.06%)	105,689 (1.48%)
Repeats[a)]	414,971 (2.24%)	167,419 (2.34%)
No annotation[a)]	16,466,792 (88.71%)	6,858,935 (95.78%)

a) Mapped to the *G. raimondii* whole genome sequence (http://cgp. genomics.org.cn/ page/species/index.jsp); b) The number of reads includes perfectly matched miRNAs and their ⩽2 nt mismatches variants.

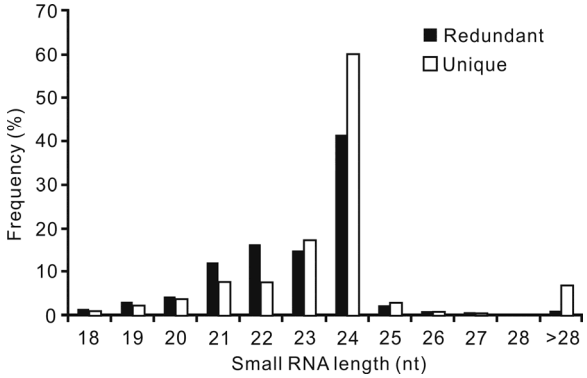

Figure 1 Length distribution and abundance of the small RNA sequences in the small RNA library of 15-DPA cotton seeds.

2.2 Identification and expression analysis of known miRNAs

Many miRNAs are evolutionarily conserved in the plant kingdom. Currently, 55 families containing 84 mature miRNAs are contained in miRBase (release 21). To identify conserved miRNAs in cotton, all processed (clean) sequences from immature seeds at 15 DPA (S15) were pooled and searched for the presence of the known miRNAs listed in miRBase. In total, 773,117 reads representing 809 unique reads matched 95 known mature miRNAs sequences, which corresponded to 95 miRNA families (Table S2). These identified miRNA families are conserved in various plant species.

Cotton miRNA families displayed significantly varied abundance that ranged from 1 to 520,801 reads. The abun-

dance of these known miRNAs, as reflected in normalized reads (RPTM), was further compared in the present study. The five miRNA families with the higher abundance were 167 (280,563 RPTM), 156/157 (62,178 RPTM), 165/166 (23,775 RPTM), 894 (10,950 RPTM) and 172 (10,882 RPTM), which account for 67.4%, 14.9%, 5.7%, 2.6% and 2.6% of the total known miRNA reads, respectively. The miRNA families 164, 168, 396-3p, 535, 2911, 3476 and 7508 also showed average abundances of more than 1,000 RPTM. In contrast, certain miRNA families were observed to be present at lower levels. The altered miRNA expression suggests that miRNA genes are differentially transcribed at this stage of seed development. Because no genome sequence is available for allotetraploid cotton (*G. hirsutum*), we also examined known miRNA precursors by mapping known miRNA sequences to *G. raimondii* genome sequences. Among the known miRNAs, a total of 126 sequences from 25 miRNA families mapped to the genome sequence (Table S3); all of the miRNA families are able to adapt hairpin structures that resemble the fold-back structures of miRNA precursors. In addition, the number of members per miRNA family ranged from 1 to 16. The largest miRNA family size identified was miR156 that consisted of 16 members, and the miR166, miR169 and miR482 families possessed 12, 11 and 11 members, respectively. However, the miR159, miR162_1, miR408, miR530, miR827_2 and miR2111 families had only one member detected in 15-DPA seeds. Sixty-four miRNA* sequences were also detected in this study (Table S3). The detection of miRNA*s represents further evidence for the existence of mature miRNAs.

To investigate whether the expression of miRNAs was different from other tissues, another small RNA library from cotton fibers at 15 DPA (F15) used in our previous study was used as a reference in this study [19]. Among 95 known miRNA families, 48 miRNA families were differentially expressed between S15 and F15 (Figure 2). Of these families, 23 miRNA families were up-regulated in S15 compared with F15, whereas 25 miRNA families showed down-regulated patterns (Table S4). The miRNAs miR828, miR1171, miR5141, miR6118-3p, miR169_1, and miR7500-3p showed lower expression in F15 but relatively high expression in immature seeds (⩾2-fold changes). For example, miR828 was specifically expressed in the cotton seed but undetectable in the fiber, which may support previous research that miR828 targets the *MYB2* mRNA and that the temporal regulation of *MYB2* expression during ovule and fiber development is most likely mediated by miR828 in allotetraploid cotton [19]. In addition, eight of 25 miRNA families were more represented in F15 than S15 (⩾2-fold change). The eight down-regulated miRNA families were miR169_2-5p, miR397, miR477, miR319, miR390, miR395, miR7513 and miR396-5p. These results suggest a possible miRNA-mediated mechanism for gene expression control in regulating seed and fiber development.

2.3 Identification and characterization of novel cotton miRNA candidates

The unannotated small RNAs of 6,858,935 unique reads were subjected to rigorous secondary structure analysis of their precursors using the mireap software developed by BGI (Shenzhen, China). Following the three criteria suggested by Meyers et al. [11], 23 novel miRNA candidates from 28 miRNA gene loci were selected (Table 2 and Table S5) and named temporarily in the form of ghr-miRs-number format, e.g., ghr-miRs001, before being submitted to obtain an official designation. Of the 23 identified novel miRNA candidates, four miRNAs (ghr-miRs005, ghr-miRs015, ghr-miRs016 and ghr-miRs017) were linked to more than one miRNA gene locus, and the remaining 19 miRNAs were encoded by just a single locus. Moreover, 16 miRNAs were 24 nt long, while the rest seven members were 21 nt long (Table 2).

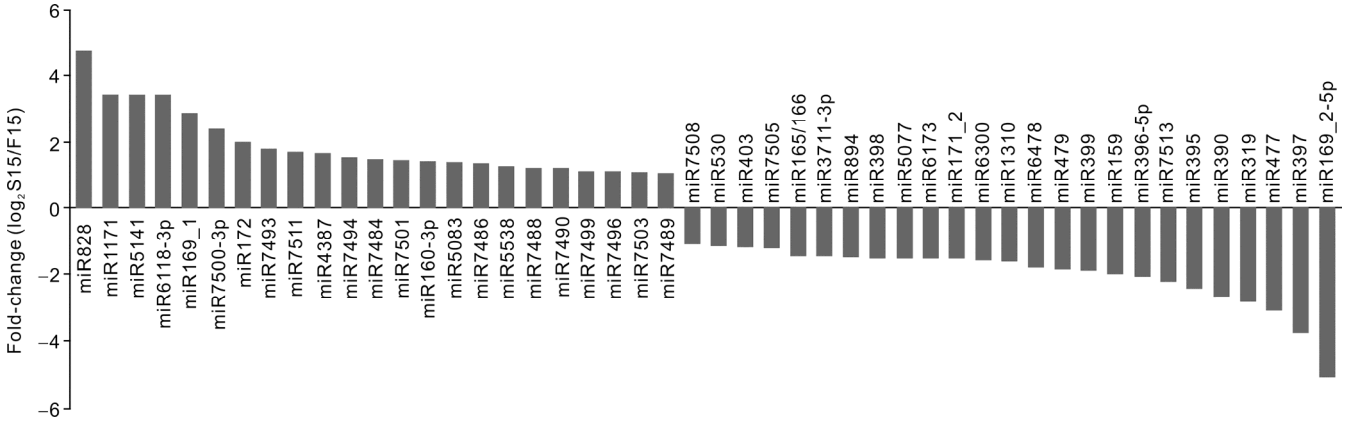

Figure 2 Differential expression of known miRNAs in 15-DPA cotton seed (S15) and fiber (F15). Small RNA library from cotton fibers at 15-DPA used in our previous study [4] was used as a reference.

Table 2 Potential novel miRNAs identified in 15-DPA cotton seeds

Name	Mature sequence (5′ to 3′)	Length (nt)	G+C (%)	Precursor number	Reads
ghr-miRs001	AUUAGGGUGGGUUUGGAUGGGCGA	24	54.17	1	7
ghr-miRs002	AGGUGAUUUGACAAACAAUGCAAG	24	37.5	1	8
ghr-miRs003	AACCAUUAGAUUUUAAUGGGUCGG	24	37.5	1	8
ghr-miRs004	CAGGAAGAGGAAGAUGAAAUA	21	38.1	1	10
ghr-miRs005	GCACUGUCAGAAAAAUGCAUUGAA	24	37.5	2	10
ghr-miRs006	AUUUCUACUGUUAAAAAUUGGCAU	24	25	1	6
ghr-miRs007	AUGUAGAUUGCCAUGUGGAUGACA	24	41.67	1	6
ghr-miRs008	AGGAGUUAGGUUGCAUUCUGCCCC	24	54.17	1	6
ghr-miRs009	UUUGUGAGGGCAAAAGAUCGA	21	42.86	1	5
ghr-miRs010	UAGGUUAGAUCAAAGAGCAAA	21	33.33	1	8
ghr-miRs011	AAAUGUUGUCACUGUGUUAGGCGG	24	45.83	1	15
ghr-miRs012	AUGUUGGAUCAAAGAGCAAAUCGG	24	41.67	1	7
ghr-miRs013	UGAGAUUGGCAUCAAUGUUCA	21	38.1	1	7
ghr-miRs014	UCAACAGGAGGACUAGUUUGC	21	47.62	1	5
ghr-miRs015	UAAAGACCAAGAACUUUAGCGGCG	24	45.83	2	10
ghr-miRs016	AUUAGCGGCGCUUUUUGAAAAAUG	24	37.5	2	11
ghr-miRs017	AAUGCUCAGGGCUUUAGUGGCGUU	24	50	3	16
ghr-miRs018	ACAGACCAAACAUACAAAUGGACC	24	41.67	1	14
ghr-miRs019	UUUACUGACGUGGCAACAAAUCGC	24	45.83	1	8
ghr-miRs020	UAUAAGGAAAGAAUUGGAUGA	21	28.57	1	7
ghr-miRs021	AACCCACUUGGGAAAUUAUGGGAU	24	41.67	1	9
ghr-miRs022	GAAAAGUACAAGGACUAUAGGCAU	24	37.5	1	9
ghr-miRs023	CAACGGUGGAGGUAUUGUGCU	21	52.38	1	38

The Minimum Folding free Energy Index (MFEI) is one of the most important parameters to distinguish miRNAs from other small RNAs. Increasing researches have shown that more than 90% of miRNA precursor sequences hold a MFEI higher than 0.85 but the values for other small RNAs are less than 0.85, suggesting that this value has become a sufficient criterion to confirm the miRNAs [12]. Thus, the MFE and MFEI were calculated for each candidate sequence (Table S5). The determined secondary structures of 28 predicted novel miRNA sequences showed ideal average values of MFE (–75.048 kcal mol^{-1}), MFEI (1.12) and GC content (37.80%). Of them, 25 out of 28 miRNA sequences have a MFEI greater than 0.85, while the MFEIs of only three RNAs are lower than 0.85. This feature agreed with previous studies [12,20]. We believe that these 28 sequences are most likely true novel miRNAs.

In contrast to the animal miRNA precursors (typically 70–80 nt), plant miRNA precursors are longer and more variable. In the present study, plant miRNA precursors vary in size from 77 to 461 nt, with an average of about 204 nt. In Figure 3 and Figure S1, the precursor sequences and secondary structures of the 23 novel miRNAs identified from our sequencing data using mireap were predicted. The predicted secondary structures of four representative novel miRNA precursors identified in 15-DPA cotton seeds are shown in Figure 3. It is obvious that the locations of these miRNAs in precursors are not unchanged with respect to RNA loops. Deep sequencing has facilitated our efforts to identify more mature miRNA variants with nucleotide variation at the 5′ and/or 3′ ends of these miRNA molecules. In addition, we also looked for sequenced miRNA* sequences; only three miRNA* sequences were found in our dataset (Table S5). The mature miRNA and miRNA* sequences in precursors are highlighted using different colors (Figure S1). The accumulated levels of miRNA* stands were observed to be much lower than those of their corresponding miRNA molecules, suggesting rapid degradation of miRNA* strands during the biogenesis of mature miRNAs [21]. In addition, the majority of the novel miRNAs identified had weak expression levels (Table S5), indicating that novel miRNAs frequently have lower levels of expression than the conserved miRNAs. The low abundance of novel miRNAs might suggest a specific role for these miRNAs during cotton seed development.

2.4 Target prediction, validation, and gene ontology analysis

It is well known that miRNAs are indeed able to suppress gene expression by promoting mRNA decay, inhibiting translation or both [22]. As the first step towards investigation of the roles for these identified miRNAs in seed development, putative target genes were predicted using the web-based tool psRNATarget program. A maximum mismatch value of 3 was used for higher prediction coverage. As shown in Table S6, a total of 205 EST (expressed sequence tag) sequences matched the 21 novel miRNAs, representing 176 unique targets with an average of eight targets per miRNA molecule. On the other hand, a total of 701 EST sequences matched the 87 known miRNAs are listed in Ta-

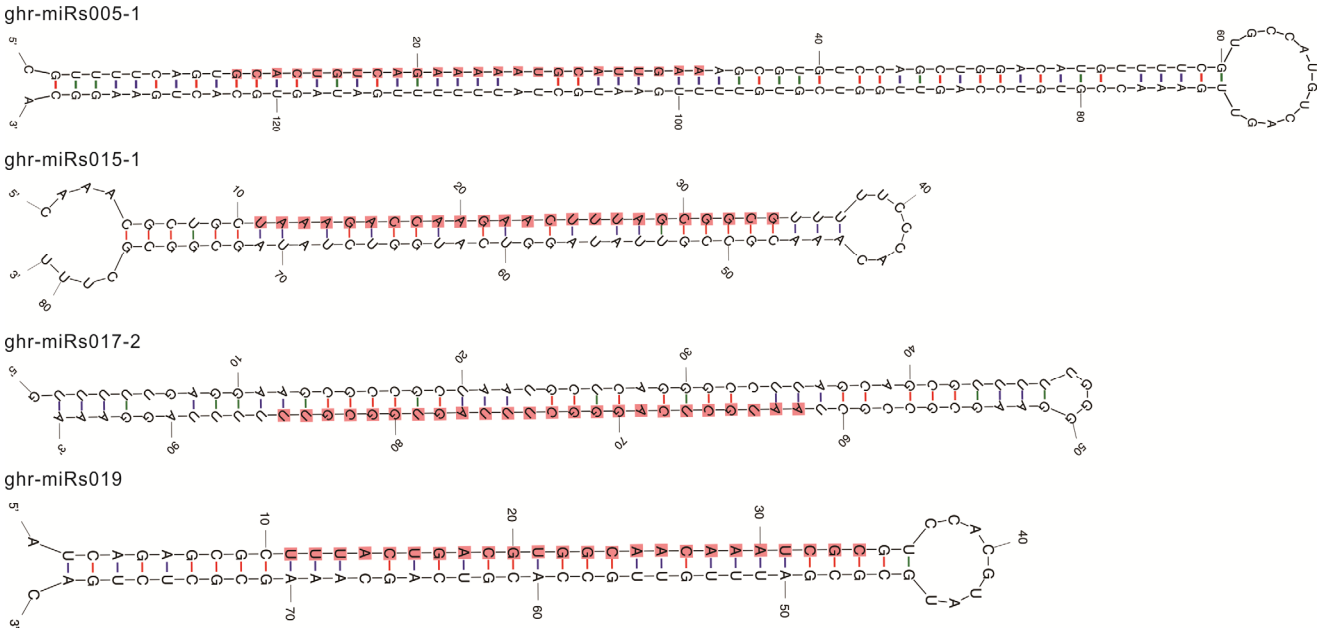

Figure 3 (color online) Predicted secondary structures of four representative novel miRNA precursors identified in 15-DPA cotton seeds.

ble S7. No target genes were found for the remaining two novel miRNAs (ghr-miRs005 and ghr-miRs021) or eight known miRNAs (miR2911, miR5077, miR5139, miR6173, miR6300, miR7486, miR7487 and miR169-3p), suggesting that these miRNAs with no predicted targets may suppress gene expression by different mechanisms including inhibiting translation. The miRNA families ghr-miRs001, ghr-miRs002, ghr-miRs003, ghr-miRs004, ghr-miRs006, ghr-miRs008, ghr-miRs009, ghr-miRs010, ghr-miRs011, ghr-miRs012, ghr-miRs013, ghr-miRs014, ghr-miRs015, ghr-miRs016, ghr-miRs020, ghr-miRs022 and ghr-miRs023 had multiple distinct targets. Ghr-miRs007, ghr-miRs017, ghr-miRs018 and ghr-miRs019 targeted only one locus. As shown in Tables S6 and S7, the predicted targets included transcription factors, enzymes, receptors, transporters and diverse cellular components involved in various biological processes, indicating the extensive roles of miRNAs in gene regulatory network. As shown in Tables S6 and S7, several miRNAs targeted transcription factors, such as the NAC and MYB transcription factors, are consistent with their functions reported previously [23,24].

Among predicted miRNA targets, cleavages of three target transcripts, as an example, were selected to be experimentally validated *in vivo* by RLM-5′ RACE. As shown in Figure 4, the cleavage products of TC272934 (squamosa promoter-binding protein-like 9, SPL9) and TC258355 (proline-rich receptor-like protein kinase 1, PEPK1) are precisely terminated at the 10th position relative to the 5′ end of the complementary regions of each associated miRNA. However, for GhmiR396-5p, the TC240832 sequence was cleaved six nucleotides upstream of the predicted site. Consistent with our previous research [4], the mRNAs of SPL9 were cleaved within the complementary region of GhmiR156. The target of miR156, SPL9, has been reported to play critical roles in the temporal control of trichome distribution in *Arabidopsis thaliana* [25].

To gain a detailed understanding of the biological significance of the target genes regulated by miRNAs in developing seeds, a total of 503 unique targets (Tables S6 and S7) were subjected to singular enrichment analysis (SEA) in AgriGO to have the gene ontology (GO) descriptions [13]. The distribution of the target genes in different GO categories including cellular components, biological processes and molecular functions is shown in Figure 5. In the biological process category, genes were highly enriched for GO terms related to cellular process, metabolic process and biological regulation. In the molecular function category, the most significant GO terms were catalytic and binding activities. For example, many *MYB* transcripts, which are predicted to be the main targets of miR828, were assigned to this category, suggesting a role for miR828 and MYB transcription factor in cotton fiber and seed development. With respect to biological processes, 251 genes primarily participate in different cellular and metabolic processes and response to stimulus, suggesting that the cotton miRNAs are involved in a broad range of physiological functions. These predicted target genes encode a broad range of proteins related to seed development and energy storage in cotton. Further analysis of their targets is needed and would provide insight into the roles these newly identified miRNAs play during seed development.

3 Conclusion

In our study, a small RNA library was constructed from

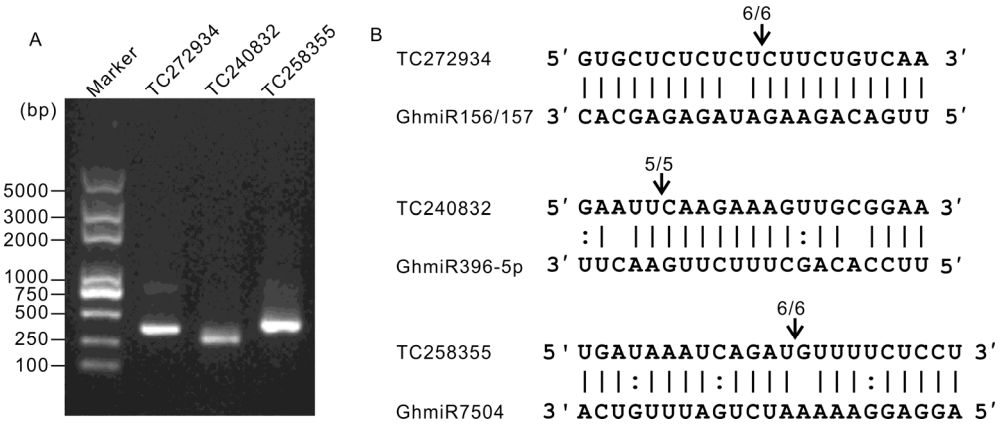

Figure 4 RLM-5′ RACE verification of predicted miRNA target genes. A, 3′ cleavage fragments of three selected targets. B, The cleavage sites of three selected genes targeted by GhmiR156/157, GhmiR396-5p and GhmiR7504. The arrow indicates a cleavage site verified by RLM-5′ RACE, with the frequency of cloned RACE products shown above the alignment.

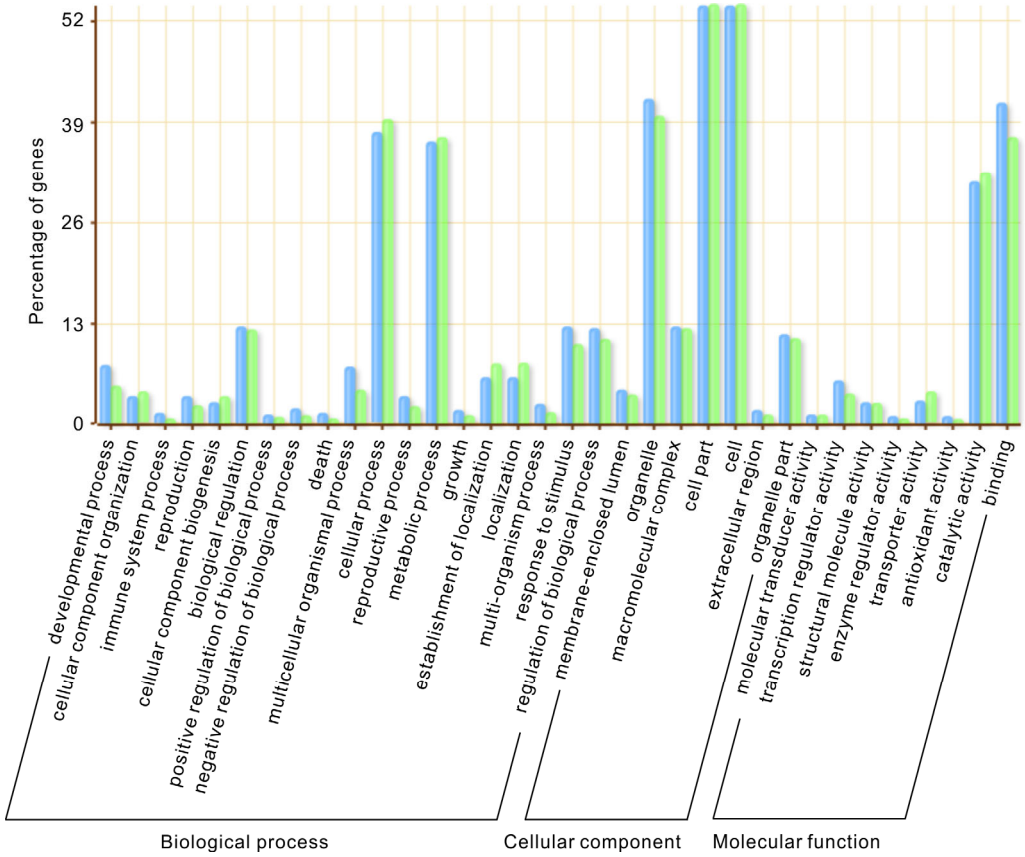

Figure 5 GO analysis of targets of known and novel miRNAs in this study. The *X*-axis shows the target gene categories. The *Y*-axis is the percentage of genes mapped by the categories, and represents the abundance of the GO categories. Blue bars indicate the enrichment of miRNA targets in the GO categories. Green bars indicate the percentage of total annotated cotton genes that map to the GO categories.

developing cotton seeds (15 DPA) for high-throughput sequencing. We identified 23 novel miRNAs that had not been reported previously in cotton. Our small RNA sequencing enlarged the cotton miRNA repertoire and confirmed the authenticity of 95 known miRNAs in developing cotton seeds. The miRNA* sequences of 20 known miRNA and three new miRNAs were also detected, providing additional evidence for the existence of miRNAs. Compared to the 15-DPA fiber miRNA dataset, 48 known miRNAs were differentially expressed between cotton seeds and fibers at 15 DPA. The putative targets for 21 novel and 87 known miRNAs were predicted. Furthermore, the respective targets of GhmiR156/157, GhmiR396-5p and GhmiR7504 were also confirmed for cleavage *in vivo* by RLM-5′ RACE. A

GO analysis of the cotton miRNA targets identified in this study provides more information about the regulatory network of miRNAs in cotton, and these results will advance our understanding of miRNA functions in seed development.

The authors declare that they have no conflict of interest.

The authors acknowledge members of the Laboratory of Molecular Biology at Tsinghua University for critical discussions. This work was supported by the National Basic Research Program of China (2010CB126003), and the National Transgenic Animals and Plants Research Project (2011ZX08005-003, 2011ZX08009-003).

1 Du SJ, Dong CJ, Zhang B, Lai TF, Du XM, Liu JY. Comparative proteomic analysis reveals differentially expressed proteins correlated with fuzz fiber initiation in diploid cotton (*Gossypium arboreum* L.). J Proteomics, 2013, 82: 113–129
2 Wang KB, Wang ZW, Li FG, Ye WW, Wang JY, Song GL, Yue Z, Cong L, Shang HH, Zhu SL, Zou CS, Li Q, Yuan YL, Lu CR, Wei HL, Gou CY, Zheng ZQ, Yin Y, Zhang XY, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu SX. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet, 2012, 44: 1098–1103
3 Wang ZM, Xue W, Dong CJ, Jin LG, Bian SM, Wang C, Wu XY, Liu JY. A comparative miRNAome analysis reveals seven fiber initiation-related and 36 novel miRNAs in developing cotton ovules. Mol Plant, 2012, 5: 889–900
4 Xue W, Wang ZM, Du MJ, Liu YD, Liu JY. Genome-wide analysis of small RNAs reveals eight fiber elongation-related and 257 novel microRNAs in elongating cotton fiber cells. BMC Genomics, 2013, 14: 629
5 Ruan YL. Boosting seed development as a new strategy to increase cotton fiber yield and quality. J Integr Plant Biol, 2013, 55: 572–575
6 Guan XY, Pang MX, Nah G, Shi XL, Ye WX, Stelly DM, Chen ZJ. Mir828 and mir858 regulate homoeologous *MYB2* gene functions in *Arabidopsis* trichome and cotton fibre development. Nat Commun, 2014, 5: 3050
7 Abdurakhmonov IY, Devor EJ, Buriev ZT, Huang L, Makamov A, Shermatov SE, Bozorov T, Kushanov FN, Mavlonov GT, Abdukarimov A. Small RNA regulation of ovule development in the cotton plant, *G. hirsutum* L. BMC Plant Biol, 2008, 8: 93
8 Pang M, Woodward AW, Agarwal V, Guan X, Ha M, Ramachandran V, Chen X, Triplett BA, Stelly DM, Chen ZJ. Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (*Gossypium hirsutum* L.). Genome Biol, 2009, 10: R122
9 Li Q, Jin X, Zhu YX. Identification and analyses of miRNA genes in allotetraploid *Gossypium hirsutum* fiber cells based on the sequenced diploid *G. raimondii* genome. J Genet Genomics, 2012, 39: 351–360
10 Yu DW, Wang YM, Xue W, Liu JY. Identification and profiling of known and novel fiber microRNAs during the secondary wall thickening stage in cotton (*Gossypium hirsutum*) via high-throughput sequencing. J Genet Genomics, 2014, 41: 553–556
11 Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Martienssen RA, Poethig RS, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu JK. Criteria for annotation of plant microRNAs. Plant Cell, 2008, 20: 3186–3190
12 Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA. Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci, 2006, 63: 246–254
13 Du Z, Zhou X, Ling Y, Zhang ZH, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res, 2010, 38: W64–W70
14 Zhao CZ, Xia H, Frazier TP, Yao YY, Bi YP, Li AQ, Li MJ, Li CS, Zhang BH, Wang XJ. Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea* L.). BMC Plant Biol, 2010, 10: 3
15 Song QX, Liu YF, Hu XY, Zhang WK, Ma B, Chen SY, Zhang JS. Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. BMC Plant Biol, 2011, 11: 5
16 Kang MM, Zhao Q, Zhu DY, Yu JJ. Characterization of microRNAs expression during maize seed development. BMC Genomics, 2012, 13: 360
17 Yi R, Zhu ZX, Hu JH, Qian Q, Dai JC, Ding Y. Identification and expression analysis of microRNAs at the grain filling stage in rice (*Oryza sativa* L.) via deep sequencing. PLoS One, 2013, 8: e57863
18 Li T, Li H, Zhang YX, Liu JY. Identification and analysis of seven H_2O_2-responsive miRNAs and 32 new miRNAs in the seedlings of rice (*Oryza sativa* L. ssp. *indica*). Nucleic Acids Res, 2011, 39: 2821–2833
19 Szittya G, Moxon S, Santos DM, Jing R, Fevereiro MP, Moulton V, Dalmay T. High-throughput sequencing of *Medicago truncatula* short RNAs identifies eight new miRNA families. BMC Genomics, 2008, 9: 593
20 Bonnet E, Wuyts J, Rouze P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics, 2004, 20: 2911–2917
21 Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. Genes Dev, 2006, 20: 3407–3425
22 Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell, 2009, 136: 215–233
23 Shang HH, Li W, Zou CS, Yuan YL. Analyses of the NAC transcription factor gene family in *Gossypium raimondii* ulbr.: chromosomal location, structure, phylogeny, and expression patterns. J Integr Plant Biol, 2013, 55: 663–676
24 Pu L, Li Q, Fan XP, Yang WC, Xue YB. The R2R3 MYB transcription factor GhMYB109 is required for cotton fiber development. Genetics, 2008, 180: 811–820
25 Yu N, Cai WJ, Wang S, Shan CM, Wang LJ, Chen XY. Temporal control of trichome distribution by microRNA156-targeted SPL genes in *Arabidopsis thaliana*. Plant Cell, 2010, 22: 2322–2335

Supporting Information

Figure S1 Secondary structure of 19 novel miRNA precursors identified in 15 DPA cotton seeds. Mature miRNAs are highlighted in red, and miRNA*s in blue, if any. Ghr-miRsxxx (*) indicates the detection of the corresponding miRNA*.

Table S1 Primers for RLM-5′ RACE mapping of miRNA cleavage sites

Table S2 The known cotton miRNA families expressed in 15-DPA cotton seeds

Table S3 Precursors of known miRNAs identified based on the *G. raimondii* genome in 15-DPA cotton seeds

Table S4 Expression levels of the known miRNAs at S15 and F15

Table S5 Precursors of novel miRNAs identified based on the *G. raimondii* genome in 15-DPA cotton seeds

Table S6 Predicted targets for the novel miRNAs in 15-DPA cotton seeds

Table S7 Predicted targets for the known miRNAs in 15-DPA cotton seeds

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.